The Ravel data set

Xavier Alameda-Pineda, Jordi Sanchez-Riera, Vojtech Franch, Johannes Wienke, Jan Cech, Kaustubh Kulkarni, Antoine Deleforge, Radu Horaud

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

The Ravel data set

Xavier Alameda-Pineda, Jordi Sanchez-Riera, Vojtěch Franc, Johannes Wienke, Jan Čech,

Kaustubh Kulkarni, Antoine Deleforge and Radu Horaud

N° 7709

August 2011

Domaine 4

ISSN 0249-6399 ISRN INRIA/RR--7709--FR+ENG

INSTITUT NATIONAL
DE RECHERCHE
EN INFORMATIQUE
ET EN AUTOMATIQUE

INRIA

centre de recherche
GRENOBLE - RHÔNE-ALPES

The Ravel data set

Xavier Alameda-Pineda*, Jordi Sanchez-Riera, Vojtěch Franc,
Johannes Wienke, Jan Čech, Kaustubh Kulkarni, Antoine Deleforge
and Radu Horaud

Domaine : Perception, cognition, interaction
Équipe-Projet PERCEPTION

Rapport de recherche n° 7709 — August 2011 — 17 pages

Abstract: In this paper, we introduce the publicly available data set *Ravel*. All scenarios were recorded using the AV robot head POPEYE, equipped with two cameras and four microphones. The recording environment was a regular meeting room enclosing all the challenges of a natural indoor scene. The acquisition setup is fully detailed as well as the design of the scenarios. Two examples of use of the data set are provided, proving the usability of the *Ravel* data set. Since the current trend is to design robots able to interact with unconstrained environments, this data set provides several scenarios to test algorithms and methods aiming to satisfy this design constraints. The data set is publicly available at the following URL: `http://ravel.humavips.eu/`

Key-words: Human robot interaction, data set, audio-visual.

This work was supported by the European project HUMAVIPS, under EU grant FP7-ICT-2009-247525.

* Corresponding author: xavier.alameda-pineda@inrialpes.fr

Centre de recherche INRIA Grenoble – Rhône-Alpes
655, avenue de l'Europe, 38334 Montbonnot Saint Ismier
Téléphone : +33 4 76 61 52 00 — Télécopie +33 4 76 61 52 52

La base de donnes Ravel

Résumé : Dans ce papier, nous introduisons l'ensemble des donnes disponibles publiquement *Ravel*. Tous les scnarios ont t enregistr en utilisant la tte robotique AV Popeye, quip de deux camras et quatre microphones. L'environnement d'enregistrement tait une salle de runion rgulire joignant tous les dfis d'une scne naturelle intrieur. La configuration d'acquisition est entirement dtaill ainsi que la conception des scnarios. Deux exemples d'utilisation de l'ensemble des donnes sont fournies, prouvant la convivialit de l'ensemble de donnes *Ravel*. Depuis la tendance actuelle est de concevoir des robots capables de interagir avec les environnements sans contrainte, cet ensemble de donnes fournit plusieurs scnarios pour tester des algorithmes et des mthodes visant satisfaire ces contraintes de conception. L'ensemble de donnes est accessible au public l'adresse suivante: `http://ravel.humavips.eu/`

Mots-clés : Interaction home-machine, base de donnes, audio-visuel.

Contents

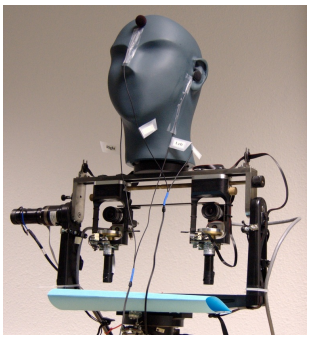


Figure 1: The POPEYE robot head was used to collect the RAVEL data set. The color-camera pair as well as two (front and left) out of four microphones are shown in the image. Four motors provide the rotational degrees of freedom and ensure the stability of the device and the repeatability of the recordings.

1 Introduction

The presence of robots in today's society is increasing. Robots are used to fulfill several tasks in, for instance, factory automation, health care, or entertainment. Nowadays, researchers and manufacturers trend to design robots able to interact with human beings in completeley unstructured and unconstrained environments. For interactive tasks, robots need to coordinate their perceptual, communicative and motor skills.

In this paper we introduce and describe in detail the publicly available data set RAVEL (robots with audio-visual interactive abilities). The data set focuses mainly on the perceptual and communicative aspects and it consists of three categories: action recognition, robot gestures and interaction. A detailed description of the categories and of the scenarios inside the categories is given below. All scenarios were recorded using an audio-visual (AV) robot head, shown in Figure 1, equipped with two cameras and four microphones, which provide a stereoscopic video stream and a quadraural audio stream per scenario.

Researchers working in multimodal human-robot interaction can benefit from RAVEL for several reasons. First of all, four microphones are used in order to be able to study the sound source separation problem; robots will face this problem when interacting with humans and/or other robots. Secondly, the simultaneous recording of stereoscopic image pairs and microphone pairs give an opportunity to test multimodal fusion methods [**?**] in the particular case of visual and auditory data. Moreover, the fact that a human-like robot head is used, makes the data appropriate to test methods intended to be implemented on humanoid robots. Finally, the scenarios are designed to study action and gesture recognition, localization of auditory and visual events, dialog handling, gender and face detection, and identity recognition. In summary, many different HRI-related applications can be tested and benchmarked by means of this data set.

The RAVEL data set is novel since it is the first data set devoted to study the human robot interactions consisting of synchronized binocular image sequences and binaural/quadraural audio tracks. The stability of the acquisition device ensures the repeata-

bility of recordings and, hence, the significance of the experiments using the data set. In addition, the scenarios were designed to benchmark algorithms aiming at different applications as described later on. To the best of our knowledge, there is no equivalent publicly available data set in terms of data quality and scenario design.

The remainder of the paper is structured as follows. Section 2 delineates the related existing data sets. Section 3 is devoted to describe the acquisition setup: the recording device, the recording environment and the characteristics of the acquired data. A detailed description of the categories and of the scenarios is given in section 4. Afterward, the data set annotation procedure is discussed (section 5). Before drawing the conclusions (section 7), some examples of usage of the RAVEL data set are given (section 6).

2 Related data sets

The RAVEL data set is at the cross-roads of several HRI-related research topics, such as robot vision, audio-visual fusion, sound source separation, dialog handling, etc. Hence, there are many public data sets related to RAVEL. These data sets are reviewd in this section and the most relevant ones are described.

Accurate recognition of human actions and gestures is of prime importance in HRI. There are two tasks in performing human actions recognition from visual data: classification of actions and segmentation of actions. There are several available databases for action recognition. KTH [**?**], Youtube Action Classification [13] and Hollywood1 [11] are data sets devoted to provide a basis for solving the action classification task. For the detection task two data sets are available: Hollywood2 [15] and Coffee and Cigarettes [22]. All these data sets provide *monocular* image sequences. In contrast, the INRIA XMAS data set [21] provides 3D visual hulls and it can be used for the classification and localization tasks. In the INRIA XMAS database, the actors perform actions in a predefined sequence and are recorded using a complex multiple camera setup that operates in a specially arranged room.

Audio-visual perception [**?**, **?**] is an useful skill for any entity willing to interact with human beings, since it provides for a spatio-temporal representation of an event. There are several existing databases for the AV research community. In particular, a strong effort has been made to produce a variety of multi-modal databases focusing on faces and speech, like the AV-TIMIT [10], GRID [7], M2VTS [19], XM2VTSDB [16], Banca [3], CUAVE [18] or MOBIO [14] databases. These databases include individual speakers (AV-TIMIT, GRID, M2VTS, MOBIO, XM2VTSDB, Banca) or both individual speakers and speaker pairs (CUAVE). All have been acquired with one close-range fixed camera and one close-range fixed microphone. Two corpora more closely related to RAVEL are the AV16.3 data set [12] and the CAVA data set [2]. Both include a range of situations. From meeting situations where speakers are seated most of the time, to motion situations, where speakers are moving most of the time. The number of speakers may vary over time. Whilst for the AV16.3 data set three fixed cameras and two fixed 8-microphone circular arrays were used, for the CAVA data set two cameras and two microphones were mounted in a person's head. Instead, RAVEL uses an active robot head equipped with far-range cameras and microphones.

Figure 2: Two views of the recording environment. The POPEYE robot is in one side of the room. As shown, the sequences were shot with and without day light providing for lighting variations. Whilst two diffuse lights were included in the setup to provide for good illumination, no devices were used to modify neither the illumination changes nor the sound characteristics of the room. Hence, the recordings are affected by all kind of audio and visual interferences and artifacts present in natural indoor scenes.

Concerning HRI data sets, [23] provides typical robotic sensors' data of a "home tour" scenario annotated using human spatial concepts; this allows to evaluate methods trying to semantically describe the geometry of an indoor scene. In [17], the authors present a new audio-visual corpus containing information of two of the modalities used by humans to communicate their emotional states; namely speech and facial expression in the form of dense dynamic 3D face geometries.

Different data sets used different devices to acquire the data, depending on the purpose. In the next section, the acquisition setup used in RAVEL, which includes the recording environment and device, is fully detailed. Furthermore, the type of recorded data is specified as well as its main properties in terms of synchronization and calibration.

3 Acquisition setup

Since the purpose of the RAVEL data set is to provide data for benchmarking methods and techniques for solving HRI challenges, two requirements have to be addressed by the setup: a robocentric collection of accurate data and a realistic recording environment. In this section this setup is described, showing that the two requisites are satisfied to a large extent. In a first stage the recording device is described. Afterward, the acquisition environment is delineated. Finally the properties of the acquired data in terms of quality, synchrony and calibration are detailed and discussed.

The POPEYE robot was designed in the framework of the POP project[1]. This robot is equipped with four microphones and two cameras providing for auditory and visual sensorial faculties. The four microphones were mounted on a dummy-head, as shown

[1] `http://perception.inrialpes.fr/POP/`

in Figure 1, designed to imitate the filtering properties associated with a real human head. Both cameras and the dummy head were mounted on a four-motor structure that provides for accurate moving capabilities: pan motion, tilt motion and camera vergence.

The POPEYE robot has several remarkable properties. First of all, since the device is alike the human being, it is possible to carry out psycho-physical studies using the data acquired with this device. Secondly, the use of the dummy head and the four microphones, allows for the comparison between using two microphones and the Head Related Transfer Function (HRTF) against using four microphones without HRTF. Also, the stability and accuracy of the motors ensure the repeatability of the experiments. Finally, the use of cameras and microphones gives to the POPEYE robot head audio-visual sensorial capabilities in one device that geometrically links all six sensors.

All sequences from the data set were recorded in a regular meeting room, shown in figure 2. Whilst two diffuse lights were included in the setup to provide for good illumination, no devices were used to modify neither the cause by sunlight nor the sound characteristics of the room. Hence, the recordings are affected by exterior illumination changes, acoustic reverberations, outside noise, and all kind of audio and visual interferences and artifacts present in unconstrained indoor scenes.

For each recorded sequence, we acquired several streams of data distributed in two groups: the *primary* data and the *secondary* data. While the first group is the data acquired using the POPEYE robot's sensors, the second group was acquired by means of devices external to the robot. The *primary* data consists of the audio and video streams captured using POPEYE. Both, left and right, cameras have a resolution of 1024×768 and two operating modes: 8-bit gray-scale images at 30 frames per second (FPS) or 16-bit YUV-color images at 15 FPS. The four Soundman OKM II Classic Solo microphones mounted on the Sennheiser MKE 2002 dummy-head were linked to the computer via the Behringer ADA8000 Ultragain Pro-8 digital external sound card sampling at 48 kHz. The *secondary* data are meant to ease the task of manual annotation for ground-truth. These data consist of one flock of birds (FoB) stream (by Ascension technology) to provide the absolute position of the actor in the scene and up to four wireless close-range microphones PYLE PRO PDWM4400 to capture the audio track of each individual actor.

Both cameras were synchronized by an external trigger controlled by software. The audio-visual synchronization was done by means of a clapping device. This device provides an event that is sharp – and hence, easy to detect – in both audio and video signals. The FoB was synchronized to the visual stream in a similar way: with a sharp event in both FoB and video signals. Regarding the visual calibration, the state-of-the-art method described in [4] uses several image-pairs to provide an accurate calibration. The audio-visual calibration is manually done by annotating the position of the microphones with respect to the cyclopean coordinate frame [8].

Following the arguments presented in the previous paragraphs it can be concluded that the setup suffices conceptual and technical validation. Hence, the sequences have an intrinsic value when used to benchmark algorithm targeting HRI applications. The next section is devoted to fully detail the recorded scenarios forming the RAVEL data set.

4 Data Set description

The RAVEL data set has three different categories of scenarios. The first one is devoted to study the recognition of actions performed by a human being. With the second category we aim to study the audio-visual recognition of gestures addressed to the robot. Finally, the third category consists of several scenarios; they are examples of human-human interaction and human-robot interaction.

4.1 Action recognition

The task of recognizing human-solo actions is the motivation behind this category; it consists of only one scenario. Twelve actors perform a set of nine actions alone and in front of the robot. There are eight male actors and four female actors. Each actor repeats the set of actions six times in different – random – order, which was promted in two screens to guide the actor. This provides for various co-articulation effects between subsequent actions. The following is a detailed list of the set of actions: (i) *stand still*, (ii) *walk*, (iii) *turn around*, (iv) *clap*, (v) *talk on the phone*, (vi) *drink*, (vii) *check watch* (analogy in [21]), (viii) *scratch head* (analogy in [21]) and (ix) *cross arms* (analogy in [21]).

4.2 Robot gestures

Learning to identify different gestures addressed to the robot is another challenge in HRI. Examples of such gestures are: waving, pointing, approaching the robot, ... This category consists of one scenario in which the actor performs six times the following set of nine gestures: (i) *wave*, (ii) *walk towards the robot*, (iii) *walk away from the robot*, (iv) *gesture for 'stop'*, (v) *gesture to 'turn around'*, (vi) *gesture for 'come here'*, (vii) *point action*, (viii) *head motion for 'yes'* and (ix) *head motion for 'no'*. In all cases, the action is accompanied by some speech corresponding to the gesture. In total, eleven actors (nine male and two female) participated in the recordings. Different English accents are present in the audio tracks which makes the speech processing challenging.

4.3 Interaction

This category contains the most interactive part of the data set, i.e. human-human as well as human-robot interaction. Each scenario consists of a natural scene in which several human beings interact with each other and with the robot. In some cases one of the actors and/or the robot act as a passive observer. This category contains six different scenarios detailed in the following. In all cases, a person emulated the robot's behaviour.

Asking for directions [AD]

In this scenario an actor asks the robot for directions to the toilets. The robot recognizes the question, performs gender identification and gives the actor the right directions to

the appropriate toilets. Six different trials (four male and two female) were performed. The transcript of this scenario is in Script 1.

Actor	(enters the scene)
Actor	Excuse me, where are the toilets?
Robot	Gentleman/Ladies are to the left/right and straight on 10 meters.
Actor	(leaves the scene)

Script 1: The script encloses the text spoken by the actor as well as by the robot in the "*Asking for directions*" scenario.

Chatting [C]

We designed this scenario to study the robot as a passive observer in a dialog. The scenario consists of two people coming into the scene and chatting for some undetermined time, before leaving. There is no fixed script – occasionally two actors speak simultaneously – and the sequences contain several actions, e.g. hand shaking, cheering, ... Five different trials were recorded.

Cocktail Party Problem [CPP]

Reviewed in [9], the Cocktail Party Problem has been matter of study for more than fifty years (see [6]). In this scenario we simulated the cocktail party effect: five actors freely interact with each other, move around, appear/disappear from the camera field of view, occlude each other and speak. There is also background music and outdoor noise. In summary this is one of the most challenging scenarios in terms of audio-visual scene analysis, action recognition, speech recognition, dialog engaging and annotation. In the second half of the sequence the robot performs some movements.

Where is Mr. Smith? [MS]

The scenario was designed to test skills such as face recognition, speech recognition and continuous dialog. An actor comes into the scene and asks for Mr. Smith. The robot forwards the actor to Mr. Smith's office. However, he is not there and when he arrives, he asks the robot if someone was looking for him. The robot replies according to what happened. The transcript for the scenario is in Script 2. Seven trials (five male and two female) were recorded to provide for gender variability.

Introducing people [IP]

This scenario involves a robot interacting with 3 people in the scene. There are two versions of this scenario: passive and active. In the passive version the camera is static, while in the active version the camera is moving to look directly at speaker's faces. Together with the *Cocktail Party Problem* scenario, they are the only exception where the robot is not static in this database.

Actor	(enters and positions him in front of the robot)
Actor	I am looking for Mr. Smith?
Robot	Yes Sir, Mr. Smith is in Room No. 22
Actor	(leaves the scene)
Mr. Smith	(enters the scene)
Mr. Smith	Hello Robot.
Robot	Hello Mr. Smith.
Robot	How can I help you?
Mr. Smith	Haven't you seen somebody looking for me?
Robot	Yes, there was a gentleman looking for you 10 minutes ago.
Mr. Smith	Thank you Bye.
Robot	You are welcome.
Mr. Smith	(leaves the scene)

Script 2: Detail of the text spoken by both actors (Actor and Mr. Smith) as well as the Robot in the "*Where is Mr. Smith?*" scenario.

In the passive version of the scenario, Actor 1 and Actor 2 interact together with the Robot and each other; Actor 3: only interacts with Actor 1 and Actor 2. The transcript of the passive version is in Script 3. In the active version, Actor 1 and Actor 2 interact with the Robot and each other; Actor 3 enters and leaves room, walking somewhere behind Actor 1 and Actor 2, not looking at the Robot. The transcript of the active version is detailed in Script 4

4.4 Background clutter

Since the RAVEL data set aims to be useful for benchmarking methods working in populated spaces, the first two categories of the database, action recognition and robot gestures, were collected with two levels of background clutter. The first level corresponds to a controlled scenario in which there are no other actors in the scene and the outdoor and indoor acoustic noise is very limited. During the recording of the scenarios under the second level of background clutter, other actors were allowed to walk around, always behind the main actor. In addition, the extra actors occasionally talked to each other; the amount of outdoor noise was not limited in this case.

4.5 Data download

The RAVEL data set is publicly available at `http://ravel.humavips.eu/` where a general description of the acquisition setup, of the data, and of the scenarios can be found. In addition to links to the data files, we provide previews for all the recorded sequences for easy browsing previous to data downloading.

Actor 1	(enters room, positions himself in front of robot and looks at robot)
Actor 1	Hello, I'm Actor 1.
Robot	Hello, I'm Nao. Nice to meet you.
Actor 2	(enters room, positions himself next to Actor 1 and looks at robot)
Robot	Excuse me for a moment.
Robot	Hello, I'm currently talking to Actor 1. Do you know Actor 1?
Actor 2	No, I don't know him.
Robot	Then let me introduce you two. What is your name?
Actor 2	Actor 2
Robot	Actor 2, this is Actor 1. Actor 1 this is Actor 2.
Actor 3	(enters room, positions himself next to Actor 1, looks at Actor 1and Actor 2)
Actor 3	Actor 1 and Actor 2, have you seen Actor 4?
Actor 2	No I'm sorry, we haven't seen her.
Actor 3	Ok, thanks. I'll have to find her myself then. Bye.
Actor 3	(leaves)
Actor 2	Actor 1, (turn heads towards robot)
Actor 1	We have to go too. Bye
Robot	Ok. See you later.

Script 3: Detail of the script of the scenario "*Introducing people - Passive*". The three people interact with the robot, although it is static.

Actor 1	(enters room, positions himself in front of robot and looks at robot)
Actor 1	Hello, I'm Actor 1.
Robot	Hello, I'm Nao. Nice to meet you.
Actor 2	(enters room, positions himself next to Actor 1 and looks at robot)
Robot	Excuse me for a moment.
Robot	(turns head towards Actor 2)
Actor 1	(turns head towards Actor 2)
Robot	Hello, I'm currently talking to Actor 1. Do you know Actor 1?
Actor 2	No, I don't know him.
Robot	Then let me introduce you two. What is your name?
Actor 2	Actor 2
Robot	Actor 2 this is Actor 1. (turns head towards Actor 1) Actor 1 this is Actor 2.
Actor 3	(enters room, walks somewhere behind Actor 1 and Actor 2, leaves room)
Actor 1	We have to go now. Bye
Robot	(turns head towards Actor 1)
Robot	Ok. See you later.

Script 4: Detail of the script of the scenario "*Introducing people - Active*". Two out of the three people interact with the robot. The latter is a moving robot.

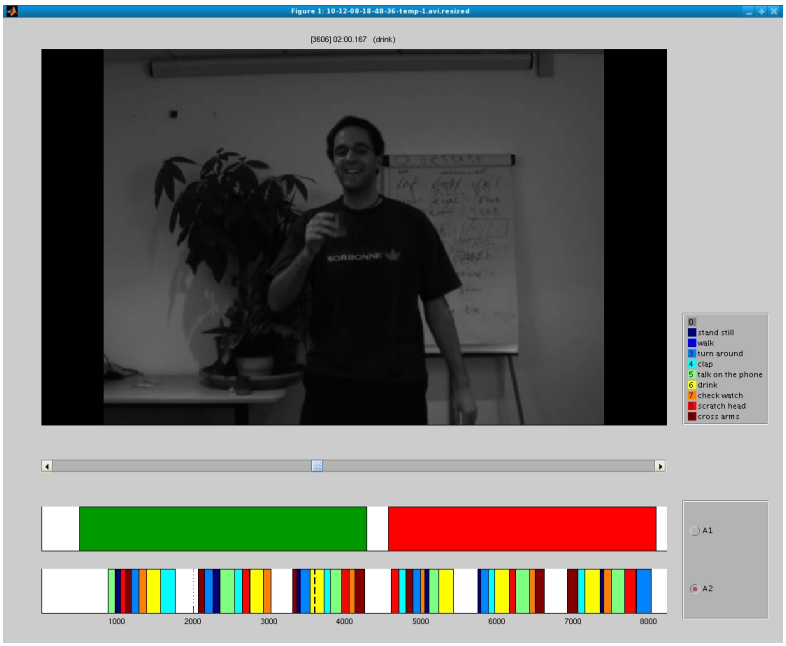


Figure 3: The annotation tool screen shot. Two time lines are shown below the image. The first one (top) is used to annotate the level of background clutter. The second one (bottom) details which action is performed at each frame.

5 Data Set annotation

Providing the ground truth is an important task when delivering a new data set; that allows to quantitatively compare the algorithms and techniques applied to the data. On one hand, the annotation for the action and gesture recognition categories is provided; this annotation is done using a classical convention, that each frame is assigned a label of the particular action. Since the played action is known only one label is assigned to each frame. Because the annotation we need is not complex a simple annotation tool was designed for this purpose in which a user labels each start and end of each action/gesture in the recordings. The output of that tool is convertible to standard formats like ELAN [**?**]. A screen shot of the annotation tool is shown in Fig. 3. On the other hand, since there are more complex scenes populated with multiple people, a localization (in 3D or in the images) of the actor performing the defined action will be provided.

6 Data exploitation examples

In order to prove the importance of the RAVEL data set, a few data exploitation examples are provided. These examples show how diverse applications can use the presented

data set. Two different examples are explained in this section: a scene flow extraction method and an event-detection algorithm based on statistical audio-visual fusion techniques.

6.1 Scene flow

Since the entire database is captured by synchronized and rectified cameras, it is possible to compute a 3D scene flow [20]. The 3D scene flow is a classical computer vision problem. It is defined as a motion field such that each reconstructed pixel for a frame has assigned a 3D position and a 3D velocity. It leads to an image correspondence problem, where one has to simultaneously find corresponding pixels between images of a stereo pair and corresponding pixels between subsequent frames.

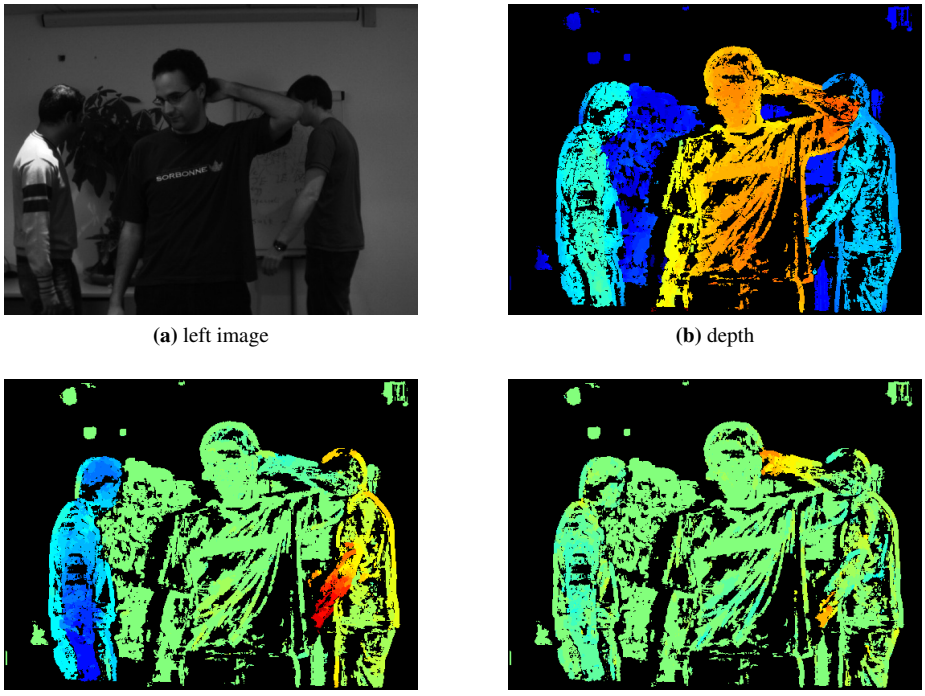
After the 3D reconstruction using the known camera calibration, these correspondences fully determine the 3D scene flow. A projection of a scene flow is shown in Fig. 4, as a disparity (or depth) map and horizontal and vertical optical flow maps. These results are computed using a recent seed growing algorithm [5]. The scene flow results can be used for further processing towards the understanding of a dynamic scene.

6.2 Audio-visual event detection

How to detect audio-visual events, i.e. events that are both heard and seen, is a topic of interest for researchers working in multimodal fusion. An entire pipeline – from the raw data to the concept of AV event – is exemplified in this section. This pipeline consists of three modules: visual processing, auditory processing and audio-visual fusion. In the following, the method is roughly described; interested readers can find a more detailed explanation in [1].

To extract visual features, interest Harris points are computed and filtered to keep those image locations related to motion. Stereo-matching is performed to later on reconstruct the points in the 3D space. The audio features are the so called Interaural Time Differences (ITD), measuring the different of time arrival between the two microphones. To fuse the two modalities, the geometrical properties of the recording device, i.e. the microphone positions in the 3D space, are used to map the 3D points into the ITD space. A modified version of the EM algorithm uses the mapped visual features to supervise the fusion of audio features into visual clusters. These clusters are back-projected to the 3D space providing localization of audio-visual events.

This example of data exploitation was applied onto the *CPP* sequence of the RAVEL data set. Figure 5 shows the results of the method in nine frames of the sequence. In this sequence the AV events are people in an informal social gathering. Although the method has some false positives, it correctly detects and localizes 26 objects out of 33 (78.8%).



(a) left image **(b)** depth

(c) horizontal motion **(d)** vertical motion

Figure 4: Scene Flow. The results are color coded. For disparity map (b), warmer colors are closer to the camera. For optical flow maps (c) and (d), green color stands for zero motion, while colder colors correspond to right and up motion respectively, warmer colors the opposite direction. Black color stands for unassigned disparity or optical flow.

Figure 5: A sample of the AV events detected in the *CPP* sequence of the RAVEL data set. The ellipses correspond to the localization of the events in the image plane. The method correctly detects and localizes 26 objects out of 33 (78.8%).

7 Conclusion & future work

The RAVEL data set is presented in this paper. Consisting of binocular and quadraural sequences, this data set embodies several scenarios designed to study different HRI applications. The data set is important because of two reasons. On one hand the statility and characteristics of the acquisition device ensure the quality of the recorded data and the repeatability of the experiments. On the other hand, the amount of data is enough to evaluate the relevance of the contents in order to improve the design of future HRI data sets.

The acquisition setup (environment and device) is fully detailed. Technical specifications of the recorded streams (data) are provided. The calibration and synchronization procedures, both visual and audio-visual, are described. Moreover, the scenarios are detailed; their scripts are provided, if applicable. The recorded scenarios are distributed in three categories representing different groups of applications: action recognition, robot gesture and interaction. Furthermore, the data set annotation is also described. Last but not least, two examples of data exploitation are given: scene flow extraction and audio-visual event detection. These prove the usability of the RAVEL data set.

Depending on the application, this work could be improved by providing more annotations on the data set such as: 3D/image locations for the actors, speaker activity, interaction description, ... In addition, regarding the interactive scenarios, the annotation of the interactions occurring, their time boundaries, who is involved and what kind of interaction is it, will be also provided. The present is, nevertheless, a complete data set in terms of technical specifications, since the full (visual and audio-visual) calibration and synchronization data is provided.

In summary, the paper details the RAVEL data set, its technical characteristics, its design and proves its usability. Hence, researchers working in multimodal human-robot interaction can benefit from the RAVEL data set.

References

[1] X. Alameda-Pineda, V. Khalidov, R. Horaud, and F. Forbes. Finding audio-visual events in informal social gatherings. In *To appear, ICMI*, 2011.

[2] E. Arnaud, H. Christensen, Y.-C. Lu, J. Barker, V. Khalidov, M. Hansard, B. Holveck, H. Mathieu, R. Narasimha, E. Taillant, F. Forbes, and R. P. Horaud. The CAVA corpus: synchronised stereoscopic and binaural datasets with head movements. In *ICMI*, pages 109–116, 2008. `http://perception.inrialpes.fr/CAVA_Dataset/`.

[3] E. Bailly-baillire, S. Bengio, F. Bimbot, M. Hamouz, J. Kittler, J. Mariéthoz, J. Matas, K. Messer, F. Porée, and B. Ruiz. The BANCA database and evaluation protocol. In *ICAVBPA*, pages 625–638. Springer-Verlag, 2003.

[4] J.-Y. Bouguet. Camera calibration toolbox for Matlab, 2008. `http://www.vision.caltech.edu/bouguetj/calib\_doc/`.

[5] J. Cech, J. Sanchez-Riera, and R. P. Horaud. Scene flow estimation by growing correspondence seeds. In *CVPR*, 2011.

[6] E. C. Cherry. Some experiments on the recognition of speech, with one and with two ears. *JASA*, 25(5):975–979, 1953.

[7] M. Cooke, J. Barker, S. Cunningham, and X. Shao. An audio-visual corpus for speech perception and automatic speech recognition (l). *Speech Communication*, 49(5):384–401, 2007.

[8] M. Hansard and R. P. Horaud. Cyclopean geometry of binocular vision. *JOSA*, 25(9):2357–2369, September 2008.

[9] S. Haykin and Z. Chen. The cocktail party problem. *NeCo*, 17:1875–1902, September 2005.

[10] T. J. Hazen, K. Saenko, C.-H. La, and J. R. Glass. A segment-based audio-visual speech recognizer: data collection, development, and initial experiments. In *ICMI*, ICMI '04, pages 235–242, New York, NY, USA, 2004. ACM.

[11] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld. Learning realistic human actions from movies. *CVPR*, 2008.

[12] G. Lathoud, J. marc Odobez, and D. Gatica-perez. AV16.3: an audio-visual corpus for speaker localization and tracking. In *2004 MLMI Workshop, S. Bengio and H. Bourlard Eds*. Springer Verlag, 2005.

[13] J. Liu, J. Luo, and M. Shah. Recognizing realistic actions from videos "in the wild". *CVPR*, 2009.

[14] S. Marcel, C. McCool, P. Matejka, T. Ahonen, and J. Cernocky. Mobile biometry (MOBIO) face and speaker verification evaluation. Idiap-RR Idiap-RR-09-2010, Idiap, rue Marconi 19, 5 2010.

[15] M. Marszalek, I. Laptev, and C. Schmid. Actions in context. *CVPR*, 2009.

[16] K. Messer, J. Matas, J. Kittler, and K. Jonsson. XM2VTSDB: The extended M2VTS database. In *ICAVBPA*, pages 72–77, 1999.

[17] Y. Mohammad, Y. Xu, K. Matsumura, and T. Nishida. The h3r explanation corpus human-human and base human-robot interaction dataset. In *ISSNIP*, pages 201 –206, dec. 2008.

[18] E. K. Patterson, S. Gurbuz, Z. Tufekci, and J. N. Gowdy. CUAVE: A new audio-visual database for multimodal human-computer interface research. In *ICASSP*, pages 2017–2020, 2002.

[19] S. Pigeon. M2vts database, 1996. `http://www.tele.ucl.ac.be/PROJECTS/M2VTS/`.

[20] S. Vedula, S. Baker, P. Rander, R. Collins, and T. Kanade. Three-dimensional scene flow. *IEEE Trans. on PAMI*, 27(3), 2005.

[21] D. Weinland, R. Ronfard, and E. Boyer. Free viewpoint action recognition using motion history volumes. *CVIU*, 104(2):249–257, November 2006. `http://4drepository.inrialpes.fr/public/viewgroup/6`.

[22] G. Willems, J. H. Becker, and T. Tuytelaars. Exemplar-based action recognition in video. In *BMVC*, 2009.

[23] Z. Zivkovic, O. Booij, B. Krose, E. Topp, and H. Christensen. From sensors to human spatial concepts: An annotated data set. *Robotics, IEEE Transactions on*, 24(2):501 –505, april 2008.

Centre de recherche INRIA Grenoble – Rhône-Alpes
655, avenue de l'Europe - 38334 Montbonnot Saint-Ismier (France)

Centre de recherche INRIA Bordeaux – Sud Ouest : Domaine Universitaire - 351, cours de la Libération - 33405 Talence Cedex
Centre de recherche INRIA Lille – Nord Europe : Parc Scientifique de la Haute Borne - 40, avenue Halley - 59650 Villeneuve d'Ascq
Centre de recherche INRIA Nancy – Grand Est : LORIA, Technopôle de Nancy-Brabois - Campus scientifique
615, rue du Jardin Botanique - BP 101 - 54602 Villers-lès-Nancy Cedex
Centre de recherche INRIA Paris – Rocquencourt : Domaine de Voluceau - Rocquencourt - BP 105 - 78153 Le Chesnay Cedex
Centre de recherche INRIA Rennes – Bretagne Atlantique : IRISA, Campus universitaire de Beaulieu - 35042 Rennes Cedex
Centre de recherche INRIA Saclay – Île-de-France : Parc Orsay Université - ZAC des Vignes : 4, rue Jacques Monod - 91893 Orsay Cedex
Centre de recherche INRIA Sophia Antipolis – Méditerranée : 2004, route des Lucioles - BP 93 - 06902 Sophia Antipolis Cedex

Éditeur
INRIA - Domaine de Voluceau - Rocquencourt, BP 105 - 78153 Le Chesnay Cedex (France)
http://www.inria.fr
ISSN 0249-6399